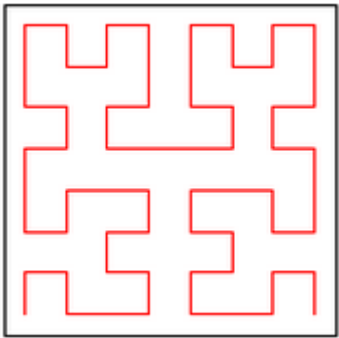


本周周报

解聪 11/4/2013-11/10/2013

本周工作

方块内布局调整。
由于之前的一个主要问题是马赛克图中的每一个方块的模式太分散，不利于识别。所以本周重点解决这个问题。之前的每一个方块内不同行业的布局采用了 Hilbert Curve 保证相似行业始终靠在一起。



问题在于各行业的企业数不一样，每个行业占据一小格会导致最终形成的方块中很稀疏，连不成大片的人眼可识别的图像模式。
因此可以使企业数较多的行业占据在一个方块中占据更多的格子。先按照企业数计算好该行业占据的格子数。

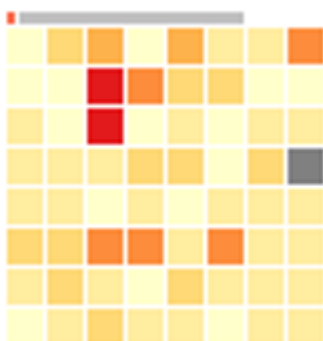


再按照 Hilbert Curve 顺序排下来，仍然可以保证相似的行业靠在一起。改进后效果如上。灰色小格代表该行业

没有企业落在这一方块。

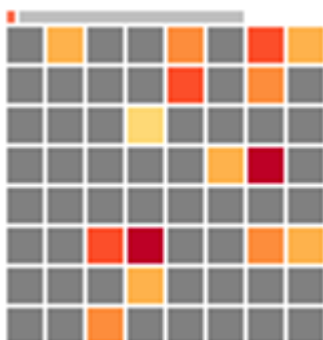
可视分析的初步探索

针对某一个像素块进行分析，比如方块 A:



该块对应的图中的 7 行, 6 列。对应的属性为: 广告>100, 科研<=0, 利润<-1000.

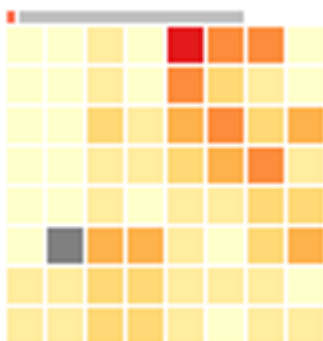
如果我们使用简单数据分析的方法, (欧式距离或 CCA 相似性) 判断其具有互补性质的块。我们可能找到了下面的块 B 与其差距最大:



该块位于第 1 行, 第 5 列, 对应的属性为广告>100, 科研<=0, 利润介于 1-99

因为这一块大部分是空的, 所以就默认其他块互补了。但事实上互补的原因仅仅是块中的总体企业数目少导致很多行业是空的, 而不是我们想要的行业模式的互补。

如果仔细观察各个行业的模式, 我们再仔细看可以发现下面的块 C:



该块对应图中 3 行 2 列。对应的属性为科研>1000, 广告<=0, 利润位于 1000-9999

再将这 AC 两块放在一起, 注意蓝色框中的区域:



再仔细对比两者的对应属性：

	科研	广告	利润
A	<=0	>100	<-1000
C	>1000	<=0	>1000

发现不论是其属性，还是其行业模式都具有非常显著的不相似性。但是当我们考虑整体时，我们就把黑色框中的行业也考虑进去了，考虑所有行业得到的效果反而不好。因此我们可以划选蓝色区域（采用之前实现的笔刷等交互），选取感兴趣区域重新计算其互补性，就可以得到比原先更优的结果。

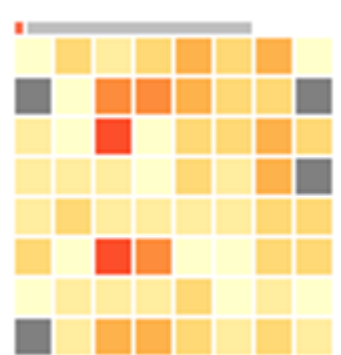
左边蓝色框中的深色的行业代码是：以 13,14,15,16,63 开头，查了一下发现 13 是食用油蜂蜜，糖类食品，14 是甜点糖果面包饼干牛奶调料等食品，15 开头是啤酒饮料等，16 开头是烟草，63 开头是供水。

右边深色的行业代码有：36,37,39,40,41 开头。36 是医学仪器，勘探仪器，交通设施，专用设备等大型仪器。37 是汽车机器修理，配件等。39 是空调，五金，洗衣机等各种大型电器。40 是通讯，雷达，监控等电子设备。41 是各种不知用途的工业机器。

对比明显发现这是左边是日常食用的轻工业，右边是专用的重工业。的确是存在很大的不同，具有互补性。

至于为什么黑色框中的行业没有太明显的差别，可能是这些行业的特征不是很明显，至少广告科科研方面各个企业分布均匀。

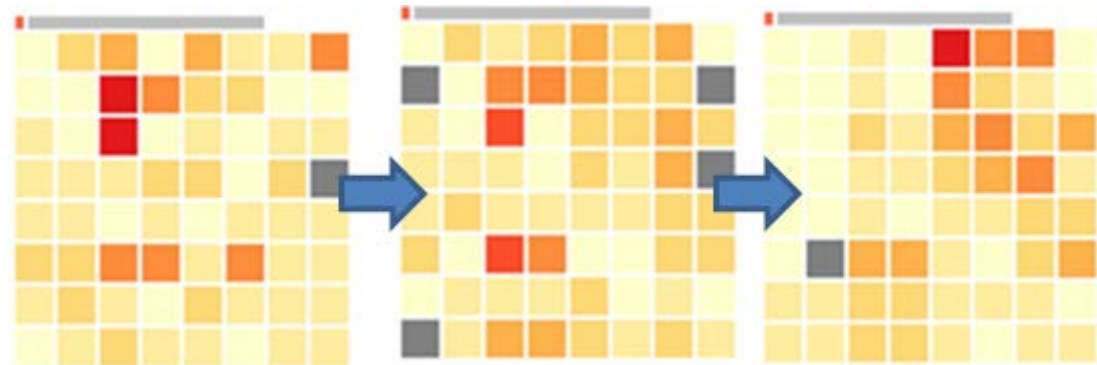
不仅如此，我还发现了一个中间状态的方块 D：



位于第 4 行，第 3 列。其属性列在下表：

	科研	广告	利润
A	<=0	>100	<-1000
C	>1000	<=0	>1000
D	1-999	>100	100-999

ACD 放在一起进行对比



下周工作

主题报告

在本周的基础上继续深入探索不同行业模式。